Efficient Model-Based Reinforcement Learning
Through Optimistic Policy Search and Planning

Sebastian Curi*, Felix Berkenkamp*, Andreas Krause

ETH zürich BOSCH

Paper Code

tldr: H-UCRL, a practical algorithm for efficient optimistic exploration in deep RL

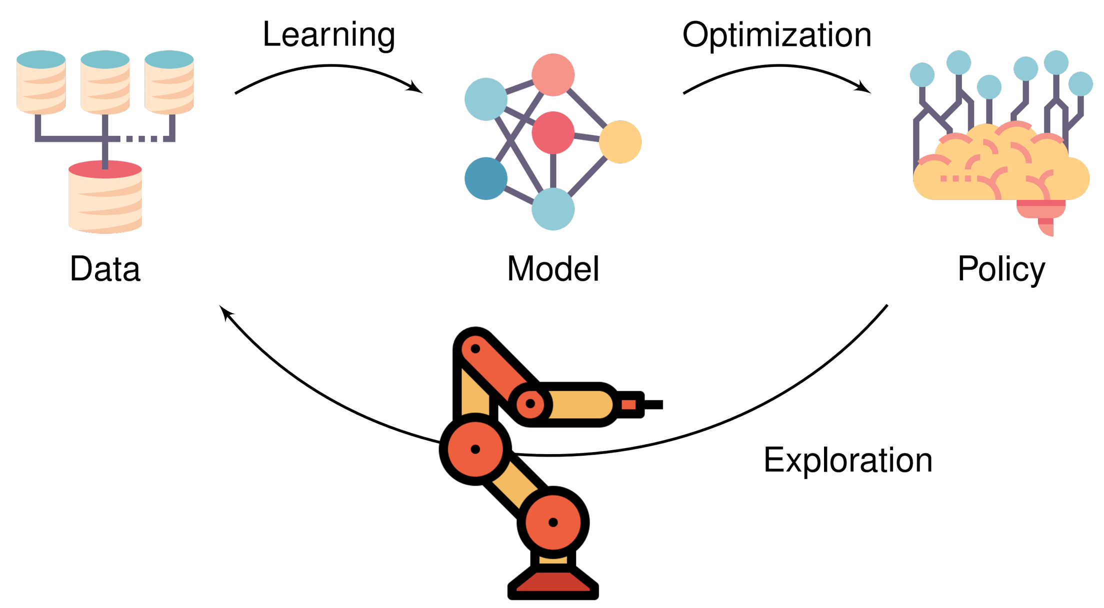
Problem Setting

Definition (Performance of policy π on model $\tilde{f}$):

$$J(\tilde{f}, \pi) = \mathbb{E}_{\omega_{0:N-1}}\left[\sum_{n=0}^{N} r(\tilde{x}_n, \pi(\tilde{x}_n)) \,\middle|\, x_0\right], \quad \text{s.t. } \tilde{x}_{n+1} = \tilde{f}(\tilde{x}_n, \pi(\tilde{x}_n)) + \omega_n.$$

Objective (Maximize performance on *true* system f):

$$\pi^* = \arg\max_{\pi \in \Pi} J(f, \pi).$$



(i) How to learn a model? Which kind of model?

(ii) How to solve the planning problem?

(iii) With which policy to collect data?

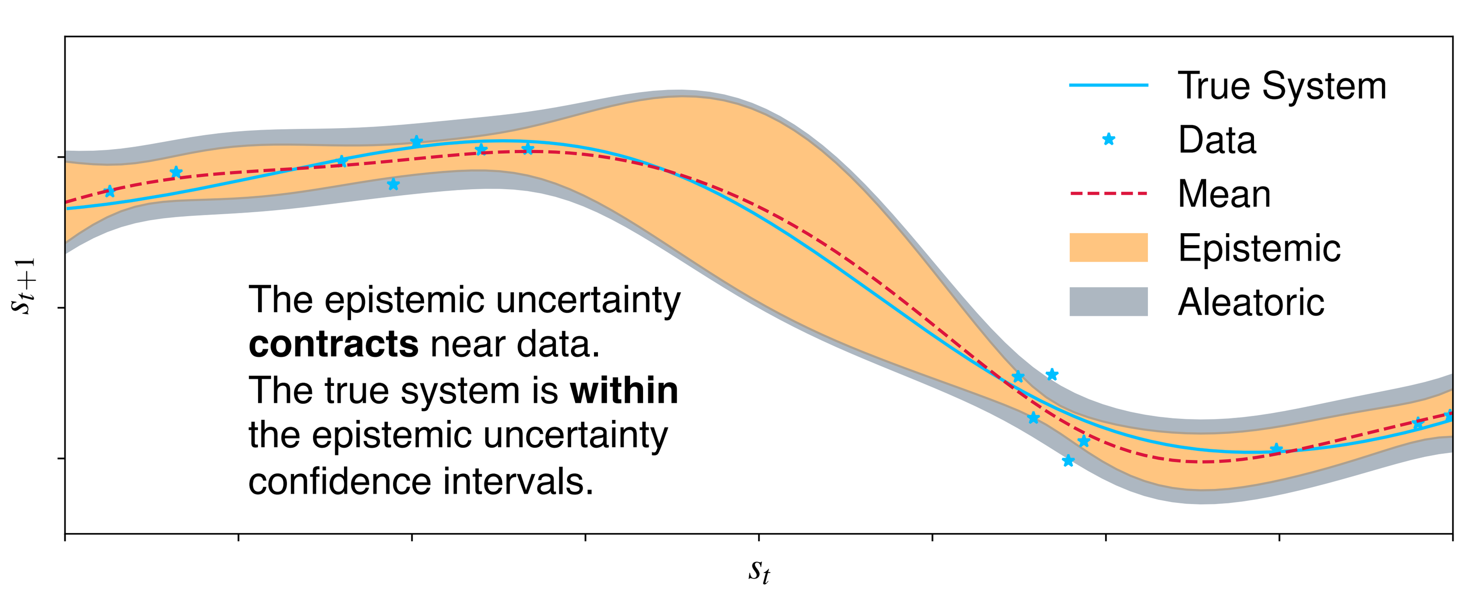
Model Learning:
Epistemic vs. Aleatoric Uncertainty

- *Aleatoric*: inherent stochasticity from the system (e.g. sensor noise).
- *Epistemic*: data scarcity (e.g. unknown weight of the tool we want to grab).

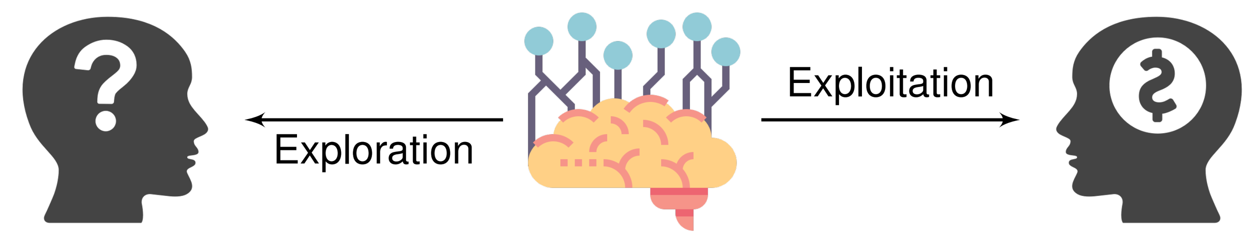
Definition (Set of Plausible Models) $\mathcal{M}_t = \{\tilde{f}, |\tilde{f} - \mu_t| \leq \beta_t \sigma_t\}$

Assumption (Well-Calibrated Models) $f \in \mathcal{M}_t \quad \forall t = 0, 1, \dots$

- GP models are calibrated under certain conditions (Srinivas et al., 2010).
- Bayesian NN models can be recalibrated empirically (Malik et al., 2019).



Data Collection:
Exploration vs. Exploitation



We want to execute a policy that:

- *Exploits* the model to maximize the performance.
- *Explores* the environment to reduce the epistemic uncertainty.

Name	Algorithm	Exploration	Implementation
Greedy	$\max_\pi \mathbb{E}_{\tilde{f}_t} J(\pi; \tilde{f}_t)$	✗	✓
UCRL	$\max_\pi \max_{\tilde{f} \in \mathcal{M}_t} J(\pi; \tilde{f})$	✓	✗
PSRL	$\max_\pi J(\pi; \tilde{f}_t),\ \tilde{f}_t \sim \mathcal{M}_t$	✓	?

- **Greedy** exploitation does *not explore* enough. Many *practical* algorithms solve such problem (e.g., PILCO (Deisenroth & Rasmussen, 2011) and PETS (Chua et al., 2018)).
- **UCRL** (Jaksch et al. 2010) instantiates the *Optimism in the Face of Uncertainty* principle and *explores efficiently*. The optimization over policies and models is *intractable*.
- **PSRL** (Osband et al. 2013) instantiates *OFU* through stochasticity but requires samples from an *exact posterior*. When the model is sampled using *approximate* methods, the exploration guarantees are lost.

H-UCRL: A reduction from UCRL to greedy

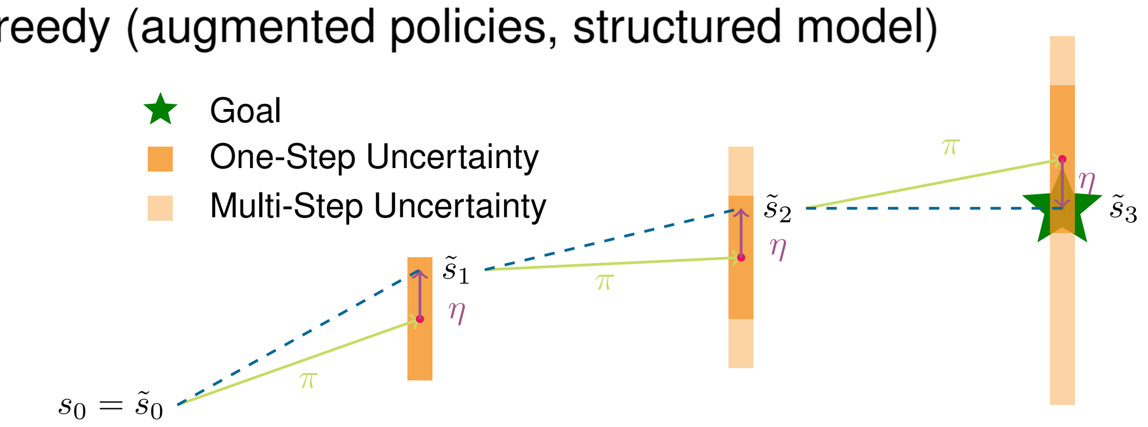
- We **reparameterize** the set of plausible models $\mathcal{M}_t$ introducing an auxiliary function η.
- H-UCRL is optimistic by **hallucinating control** over η.
- The resulting planning problem is a **tractable greedy** exploitation problem.

$$\text{UCRL} = \max_\pi \max_{\tilde{f}} J(\pi; \tilde{f}) \quad \text{s.t. } |\tilde{f} - \mu_t| \leq \beta_t \sigma_t$$

$$\text{H-UCRL} = \max_\pi \max_\eta J(\pi; \tilde{f}) \quad \text{s.t. } \tilde{f} = \mu_t + \beta_t \sigma_t \eta, \quad \eta : \mathcal{S} \to [-1, 1]$$

$$= \max_{[\pi, \eta]} J([\pi, \eta]; \tilde{f}) \quad \text{s.t. } \tilde{f} = \mu_t + \beta_t \sigma_t \eta, \quad \eta : \mathcal{S} \to [-1, 1]$$

$$= \text{Greedy (augmented policies, structured model)}$$



H-UCRL plans by:

- Using the true policy to select the next-state distribution, alike to Greedy.
- Hallucinating control to select the next state *optimistically* from within the conditional next-state distribution.

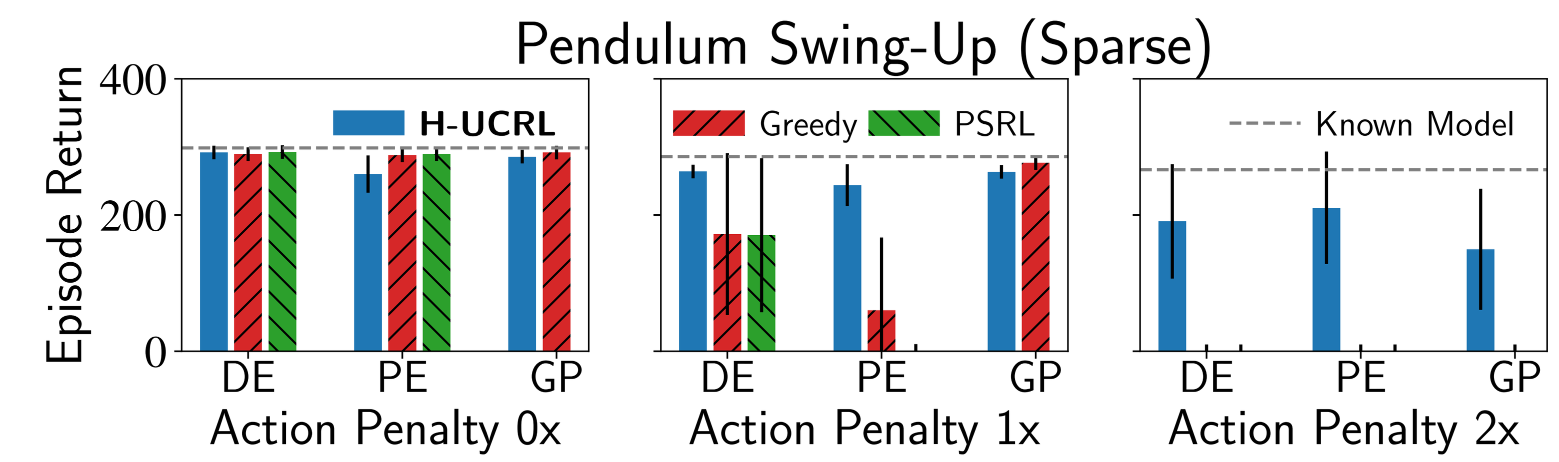
H-UCRL scales to deep NN because it does not require calibrated **multi-step** ahead predictions, but only calibrated **one-step** ahead predictions.

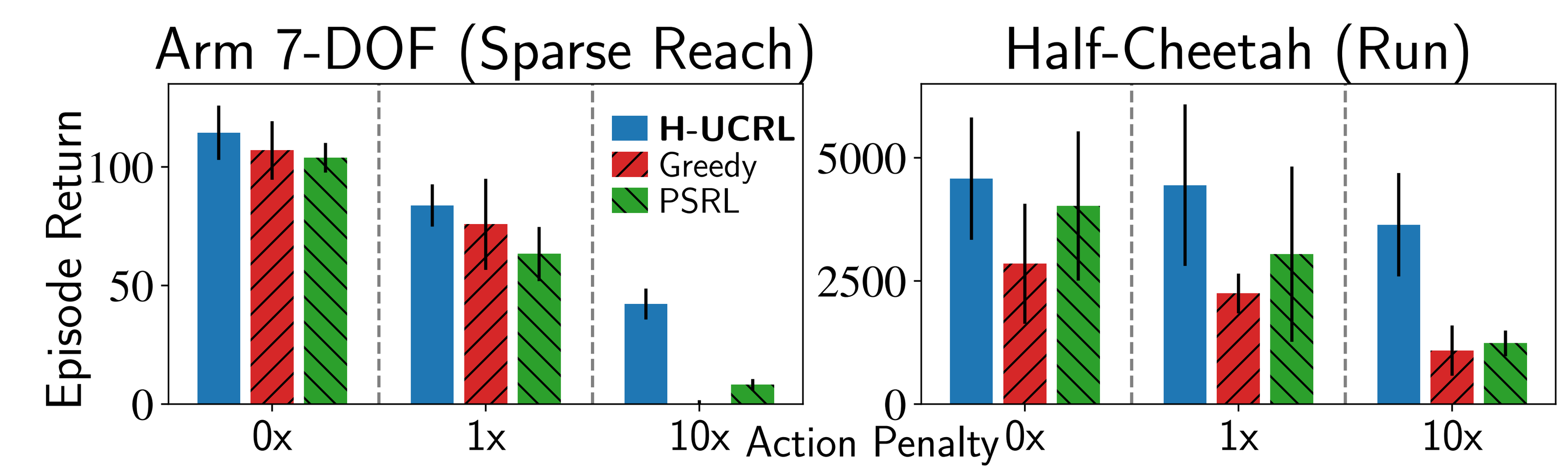
Theoretical Results (Cumulative Regret)

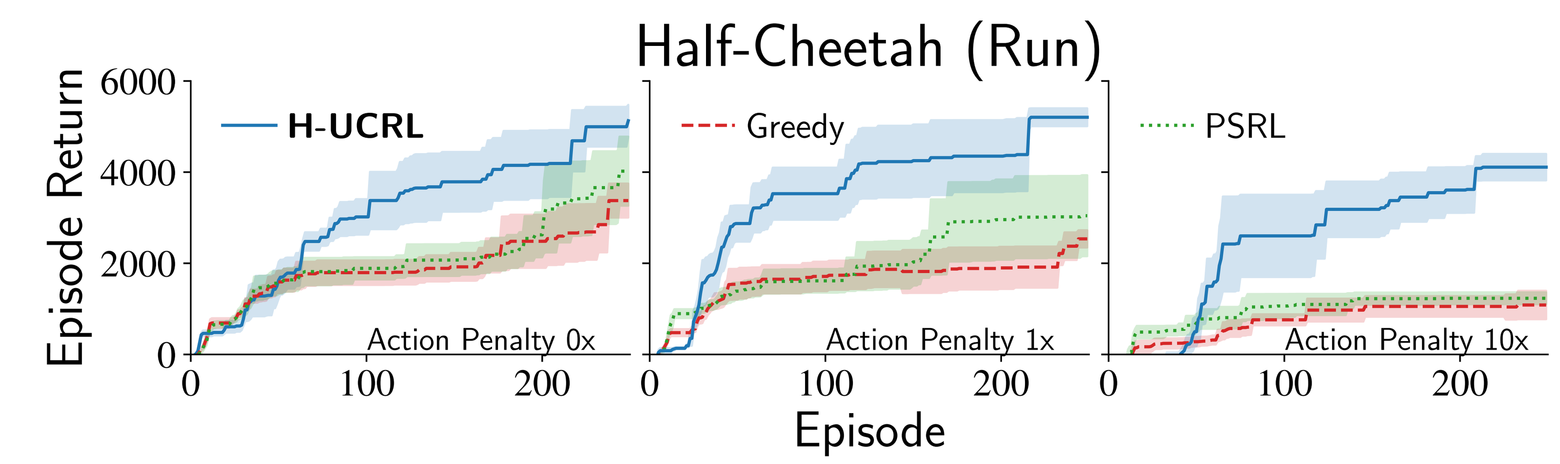
$$\text{Regret}_T = \sum_{t=1}^{T} J(f, \pi) - J(f, \pi_t^{\text{H-UCRL}}) = O(\beta_T^N \sqrt{T N^3 I_T})$$

The regret is sublinear when the maximum information gain is also sublinear. This quantity depends on the *model class* we are trying to learn. For some GP kernels, this is sublinear. The regret also depends *exponentially* on the horizon. This is the price we pay for only requiring calibrated one-step ahead predictions.

Experimental Results







- H-UCRL *outperforms* Greedy and PSRL in hard exploration problems.
- H-UCRL also outperforms Greedy and PSRL in terms of *learning speed*.

References

Deisenroth, M., & Rasmussen, C. E. (2011). PILCO: A model-based and data-efficient approach to policy search. ICML.

Chua, K., Calandra, R., McAllister, R., & Levine, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. NeuRIPS.

Jaksch, T., Ortner, R., & Auer, P. (2010). Near-optimal Regret Bounds for Reinforcement Learning. JMLR.

Osband, I., Russo, D., & Van Roy, B. (2013). (More) efficient reinforcement learning via posterior sampling. NeuRIPS.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian process optimization in the bandit setting: no regret and experimental design. ICML.

Malik, A., Kuleshov, V., Song, J., Nemer, D., Seymour, H., & Ermon, S. (2019). Calibrated Model-Based Deep Reinforcement Learning. ICML.